基于 BERT 和深度主动学习的农业新闻文本分类方法

石运来[1]，崔运鹏[1*]，杜志钢[2]

（1.中国农业科学院农业信息研究所，北京 100081；2.淄博市数字农业农村发展中心，淄博 255000）

摘　要：［目的／意义］当前农业新闻分类研究中的模型训练以被动学习方式居多，普遍存在数据无法即时标注及标注成本过高的问题，对农业新闻分析工作也造成了一定阻碍。为解决该问题，运用主动学习或者深度主动学习技术从未标注数据中选择更有价值和代表性的数据进行人工标注并构建标注数据集，提升农业新闻挖掘工作效率和效果。［方法／过程］将文本分类常用的机器学习模型结合主动学习方法分析提升效果，以及使用 BERT 模型结合 3 种采样策略进行深度主动学习训练，在共 19 847 条样本的新闻爬虫语料上以筛选出农业相关新闻为目标，通过每轮增加 30 个样本标注的迭代实验进行测试。［结果／结论］实验结果表明：主动学习方法的应用对各个模型的训练过程均有明显提升。其中 BERT 模型配合判别性主动学习采样函数，具有最优的新闻文本分类效果和最低的标注数据需求。

关键词：深度学习；农业新闻；文本分类；BERT 模型；主动学习

中图分类号：TP391.1　　文献标识码：A　　文章编号：1002-1248（2022）08-0019-11

引用本文：石运来，崔运鹏，杜志钢. 基于 BERT 和深度主动学习的农业新闻文本分类方法[J]. 农业图书情报学报，2022，34(8)：19-29.

1 引　言

随着互联网和农业信息化的快速发展，网络上的包括农业主题在内的各类新闻内容数量呈现井喷式上升，并且新闻报道的作者也开始呈现多样化，除传统的政府媒外还有许多个人或社会组织，新闻在传播过程中产生了不容忽视的社会影响，而新闻的内容也覆盖了生活的方方面面。因此为了在农业领域进行舆情监测，灾害预警，产品营销等研究，需要对大量的新闻文本进行挖掘分析，找出有价值的信息。而这些研究的前提都是要先对新闻按照主题或者领域进行分类和筛选，所以产生了对新闻按主题进行分类的需求。因此新闻主题分类任务是农业文本挖掘领域的一项基本研究。

基本的新闻分类方式是使用机器学习或者深度学习模型去进行有监督的分类模型训练和预测。许丽在 TF-IDF（Term Frequency-Inverse Document Frequency,

收稿日期：2022-03-18

基金项目：国家科技图书文献中心（NSTL）文献专项任务（2021XM45）

作者简介：石运来（1996-），男，硕士，研究方向为自然语言处理、主动学习。杜志钢（1979-），男，硕士，高级工程师，研究方向为农业农村信息化、大数据分析、建模

*通信作者：崔运鹏（1972-），男，研究员，博士生导师，研究方向为农业信息技术、农业知识管理、数据挖掘技术研究。Email：cuiyunpeng@caas.cn

词频 - 逆文档频率）文本表示的基础上使用加权朴素贝叶斯模型构建了新闻文本分类算法[1]，提升了新闻文本分类效果。但是朴素贝叶斯分类算法由于其数据必须满足贝叶斯朴素假设，因此面对超大规模文本数据时候模型性能会出现较明显的下降。郭文强基于 SVM 实现了新冠疫情虚假新闻检测[2]，比较了对虚假新闻检测中 4 种核函数的精准度，发现线性核函数分类器作为信息检测模型成果最优。田沛霖使用了 CNN-BiGRU 神经网络模型进行了新闻分类[3]，进一步提高了算法的准确率和泛化性。可见随着更复杂的模型的不断应用，新闻文本分类的效果也在不断提高。

由于新闻数量庞大且在不断增长，故训练主题分类模型需要人工标注，耗费了大量人力和时间。而主动学习（Active Learning，AL）技术是一种通过自动选择数据的标注和训练顺序可高效准确完成机器学习任务的一项技术。它假设数据的收集相对容易，但标记成本高，这符合许多文本、视觉和语音识别任务中的实际情况。它解决了在迭代式训练流程中的一个重要问题，即如果因为标注成本和项目时间等多方面的限制条件，在整个训练流程中只能在所有未标注数据中选择有限的样本子集经过人工标注后作为训练集进行模型训练，那么选择哪些样本能使得本轮模型迭代中测试准确率的最大提升？对应地在主动学习方法中有各种采样函数负责实现不同场景下最有价值数据的筛选。最流行的主动学习方式是基于池的采样[4]，它假设有一个小的标记数据集 L，并访问一个大的未标记数据集 U，每次需要从 U 中选择下一批要标记的样本。在迭代过程的每一步，主动学习算法使用 L 和 U 中的信息来从 U 中选择要标记的最佳样本 x。然后将 x 标注后添加到 L 中，这个过程重复直到我们达到所需的样本数量或分类精度。

利用主动学习方法应用到文本分类领域获得了学者们的广泛关注。黄永毅将主动学习方法应用到 SVM 支持向量机模型[5]，把新闻文本进行了财经、军事、体育、历史、科技 5 个主题的分类训练，有效地减少了样本分布不均衡对模型性能的影响。邱宁佳[6]利用密度采样的核心集主动学习算法对 SVD-CNN 深度模型进行训练，利用样本间的相似度将样本进行聚类，并在每一个聚类簇中，按照设定的规则选择最具有价值的样本进行人工标注，减少人工标注的工作量，出色完成了弹幕文本分类任务。这些在文本分类任务中和主动学习方法配合的是浅层机器学习模型或者轻量级神经网络模型。

自然语言处理（NLP）领域最新的一个重大发展是引入了预先训练过的深度文本模型，显著提高了许多 NLP 任务的最优表现。一个突出的例子是 BERT 模型[7]，它自出现以来就受到了 NLP 研究界的广泛关注。BERT 预训练模型是谷歌公司在 2018 年提出的。在 BERT 模型中使用了双向 Transformer 编码器，使得模型能够充分获取输入文本中的语义信息。然而，使用主动学习与深度预训练模型（特别是 BERT 模型）相结合的文本分类方法，迄今为止都少有相关研究。

首先，考虑到预训模型的特性，尽管预期这些模型即使使用少量的训练数据也能产生足够的性能，但目前尚不清楚已有的主动学习方法是否有效以及能在多大程度上进一步提高其分类性能。此外，最近的深度主动学习策略，如核心集[8]和深度贝叶斯方法[9]，都是在视觉领域的卷积神经网络任务模型上开发的。这些策略在 BERT 等基于 Transformer 架构的深度网络模型上的适用性尚不可知。

为了探究使用主动学习方法应用训练 BERT 模型进行新闻文本分类的方案可行性，本研究使用了自制爬虫数据集测试主动学习方法的效果，对比了对 BERT 使用深度主动学习方法和对多种机器学习模型使用主动学习方法进行训练的效果。通过多轮实验，验证了 BERT 模型的优越性并找到了和它最搭配的采样策略，发现了一种将深度主动学习技术应用在新闻文本主题分类任务中对预训练大型网络进行高效训练的可行方案。

2 研究方法

本研究的方法设计包括了主动学习流程设计，主动学习采样策略，主动学习任务模型，文本数据集构

建，实验环境和评价指标等部分。其中任务模型和主动学习采样策略相互配合共同组成了完整的主动学习方法。

2.1 主动学习流程设计

主动学习的工作原理是使用已有的采样策略从未标记样本集中选择最有价值的样本子集，通过人工标记后再对分类器进行训练[10]。这种方式中只需标记和迭代训练小部分的无标记样本就能改善任务模型质量，提升分类效果。而基于池的方法是主动学习方法中常见的一类流程，本研究也使用了这种方法。

维护一个未标注数据的集合，由选择策略在该集合中选择当前要标注和训练的数据，经过标注后再加入有标签集合作为新的训练集。其中选择策略又叫采样函数，其作用是根据预测的标签概率等信息来选择出该选择策略认为最有标注和训练价值的、对模型的预期提升最大的一批数据，详细流程如图 1 所示。

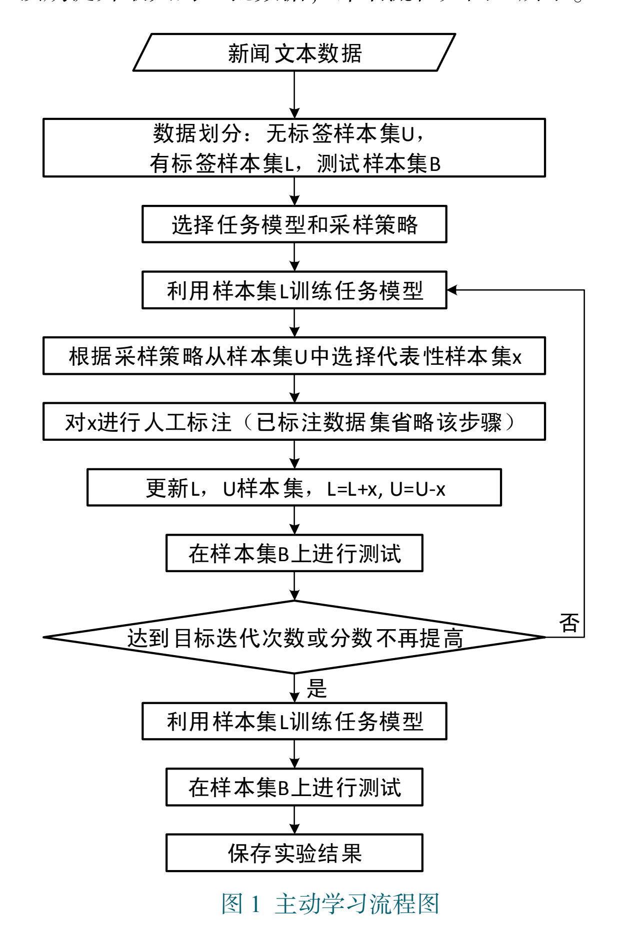


图 1 主动学习流程图

Fig.1 Flow chart of the active learning

另外，借鉴其他经验[11]直接在每轮模型的迭代训练中使用了全量训练而不是迭代训练。这种方法得到的模型精度更高，尤其是当训练深度网络的时候。

2.2 主动学习采样策略

2.2.1 最小置信度方法（Least Confidence）

该方法[12]原理是将模型在对每个数据进行预测时产生的置信度（通常是最终输出层前激活函数所获得的浮点值）作为对数据不确定性的度量。置信度越小说明模型对于这种数据越陌生，也就认为该数据越可能为模型带来更大的提升。根据置信度对未标记的样本进行升序排序，并根据每轮选取量的设置选出一个数据子集经标注后作为新的训练数据，该采样策略表示为：

$$\underset{x}{\operatorname{argmin}}\,\underset{y}{max}(\hat{P}(y \mid x)) \tag{1}$$

2.2.2 深度贝叶斯采样（Deep Bayesian Active Learning, DBAL）

深度贝叶斯采样策略专用于处理超大规模的深度神经网络，具体方法是对模型多个激活层之前加入 dropout 层，这样训练以及测试时就能够通过对 dropout 层权重的后验分布进行蒙特卡洛采样（Monte-Carlo Sampling）获得类别概率的后验分布[13]。在分类问题中，通过使用蒙特卡洛积分对近似后验概率进行求解，该采样策略表示为：

$$\begin{aligned} & p(y=c \mid x, L) \\ \approx & \int p(y=c \mid x, w) q_\theta^*(w) \mathrm{d}w \\ \approx & \frac{1}{T}\sum_{t=1}^{T} \operatorname{softmax}\left(f^{\widehat{w}_t}(x)\right) \end{aligned} \tag{2}$$

其中 T 是蒙特卡洛采样次数（在测试时深度学习模型中对给定的测试集进行重复 T 次预测），其权值为 $\widehat{w}_i \sim q_\theta^*$ (w)，q_θ (w) 为 dropout 的分布结果[10]。这样即可利用最低不确定度等采样方法在每轮迭代中根据预测概率从数据中选择出数据进行训练，形成改进的主动学习方法。

2.2.3 判别性主动学习（Discriminative Active Learning, DAL）

判别性主动学习（DAL）[14]的思路为将主动学习转换为一个二元分类任务，通过选择特定样本进行标

记，使标记池和未标记池的差别最小，从而选出最能代表样本整体的训练集样本。具体地说，Ψ：$X \rightarrow X$ 是从原始输入空间到一些学习表示的映射。该方法定义了一个二值分类问题，X 作为我们的输入空间，y 作为我们的标签空间，其中 l 是在标记集中的一个样本的标签，u 是未标记集的标签，再由算法 1 即可得到选择结果。

算法 1（Discriminative Active Learning）Input:U,L, K,n {K is the total budget,n is the a mount of mini-queries}

for i = 1…do

P←TRAIN.BINARY_CLASSIFIER(U,L)

for j = 1…$\frac{K}{n}$ do

x ←$\text{argmax}_{x \in U}$P(y=u | Ψ(x))

L←L∪x

U←U \ x

end for

end for

return U,L

2.2.4 随机选择采样（Random Sampling）

随机选择采样[15]是指采样函数从未标记数据池中随机选出一批数据作为本轮新增的训练数据。在主动学习研究中，一般将其作为基线方法与其他主动学习采样策略进行比较，验证主动学习方法的有效性。

2.3 主动学习任务模型

任务模型是主动学习方法的重要组成部分，和采样策略共同构成了完整的主动学习方法。不同的采样策略对所搭配的任务模型的也有不同要求，例如本文中的最低置信度方法只需要模型能够在预测时输出置信度即可，而 DAL 方法和 DBAL 方法则需要配合含有文本嵌入表示的深度神经网络模型。

2.3.1 机器学习模型

本研究使用几种机器学习模型与 BERT 模型进行对比，包括随机森林分类器[16]（Random Forest，RF）、多项式模型朴素贝叶斯分类器[17]（Multinomial Naive Bayesian，MNB）、逻辑回归分类器[18]（Logistic Regression，LR）、梯度提升树分类器[19]（Gradient Boosting Tree，GB）、支持向量机分类器[20]（Support Vector Machine，SVM）等。这些机器学习模型的输入数据必须是数值型数据，所以先将文本数据经过分词、TF-IDF[21]向量化等操作（取语料库中频率排名前 1 000 的词语作为 TF-IDF 特征），这样每个文本样本就表示为 1 000 维的浮点型数据。

2.3.2 BERT 模型

BERT[7]模型对于文本的表示，使用了基于 Transformer[22]架构的双向嵌入表示法，并在词嵌入之外增加了句嵌入和位置嵌入，能够很好地把握全局信息以及词与所在句子的关系，很大程度上改进了原有模型，在各项 NLP 任务中均表现突出。Transformer 层是 BERT 的主要框架，由多个编码器（Encoder）和解码器（Decoder）组成[23]。Encoder 包括 4 层：第一层为多头注意力机制（Multi-Head Attention）[24]；第二层为残差网络；第三层为前馈神经网络；第四层在最后实现了归一化操作。Decoder 在 Encoder 的基础上加入 Encoder-Decoder Attention 层，实现了信息的解码和重新序列化。其中，多头注意力层是 Transformer 层的核心，其主要思想是通过计算词与词之间的关联度来调整词的权重，反映了该词与同一句话中其他词之间的联系强弱，进而反应了每个词对于所在句子的语义表达的重要程度。首先，输入序列进入 Encoder，通过线性变换得到表示目标字的矩阵、表示上下文各个字的矩阵以及表示目标字与上下文各个字的原始矩阵。然后，通过放缩点积操作得到自注意值，该值体现了当模型对一个词进行编码时，该词与输入句子的其他词的关联程度。最后，将自注意值进行拼接和线性变换，得到与模型输入的文本长度相同的输出向量，该向量含有增强语义能够提升算法整体效果。

2.3.3 BERT 模型参数

实验参数的合理设置直接影响实验结果。模型训练过程的各项初始设置与训练参数经调整后实现了较优效果，具体参数的数值设置如表 1 所示。

在对 BERT 模型或者机器学习模型进行主动学习训练时候，保持了相同的每轮新增标注数据批次大小 30，从而能对比其在每次迭代的性能差异，以及在多

个迭代间的性能提升速度。

表 1　BERT 模型参数

Table 1　Model parameters

模型	训练参数	参数值
BERT	Hidden layer number	12
	Hidden layer dimension	768
	Network parameters	110M
	Learning rate	le-6
	Max sentence length	256
	AL suggestions num	30
	AL Epochs	8
	optimizer	AdamW

在每次模型训练过程中还使用了“提前停止（Early Stopping）[25]”的技巧。当在验证集上的准确率不再上升时及时停止训练，以此来避免过拟合、不收敛等问题，并提高模型训练效率。

2.4　文本数据集构建

2.4.1　语料获取

新闻爬虫语料。通过爬虫技术，在新浪财经、新华网等中文新闻网站上分别以“农业”“农产品”“水果”“蔬菜”等关键词爬取近 6 年的新闻文本数据，经过数据清洗后共获得 19 847 条新闻数据。其中含有农业领域新闻 9 442 条，非农业领域新闻 10 405 条。农业领域新闻中包括了农产品市场、农产品价格、农业政策发布等内容。将整个数据集按照 8:1:1 的比例划分训练集、验证集和测试集。

2.4.2　数据标注

通过人工标注的方法标记每条新闻是否为农产品相关新闻。因为主动学习的过程中需要多轮查询和数据标注，所以构建了简单的自动化标注训练系统，能够方便快速地将主动学习工作流中采样策略选择的未标记数据送往标注系统，经过四位农业领域的研究生分工标注后对模型进行训练。

2.5　实验环境

实验硬件为专业深度推理服务器，配有 8 核 CPU E5-2678 V3，ECC 内存 128G，以及 4 块 NVIDIA V100 GPU，运行 Linux 操作系统。在 Python3.6 的环境下，安装了 TensorFlow、Pytorch、Keras 等深度学习库。

2.6　评价指标

精准率是预测结果中正确预测的占比，召回率则表示样本中的正例被正确预测的占比。F_1 分数综合了精准率和召回率两个评价指标，因此更加全面，在本文中作为主要参考指标，其计算公式为两者的调和平均：F_1=2*(精准率 * 查全率)/(精准率 + 查全率)。

3　实验结果

3.1　模型选择预实验

对比 BERT 模型和不同机器学习模型在完整训练集上训练后的测试集性能。各模型的训练效果如表 2 所示。

表 2　在整个训练集上训练各个模型的效果

Table 2　The effect of training each model on the whole training set

模型	P	R	F
RF	0.792	0.776	0.784
MNB	0.823	0.825	0.824
LR	0.827	0.823	0.825
GB	0.852	0.856	0.854
SVM	0.839	0.836	0.837
BERT	0.923	0.913	0.918

对比 BERT 模型和其他模型，BERT 模型的分类效果最优，F_1 分数达到 0.918。而在其他模型中梯度提升树分类器（GB）方法最优，F_1 分数 0.854；随机森林分类器（RF）方法最差，F_1 分数 0.784；其他 3 种方法表现接近，F_1 分数在 0.824 到 0.837 之间。

但在主动学习过程中，模型的选择不只由模型的精度决定，模型的运行效率也是重要因素。在深度主动学习或者主动学习的实际应用场景的人工标注和模型训练交替进行的过程中，模型响应时间（包括了模

型训练和样本选择两个过程）过长会使标注工作在每个轮次间歇等待，浪费标注人力，降低主动学习过程的效率。重复 5 次统计在主动学习过程 0~20 轮次中各个模型响应时间并平均，结果如表 3 所示。

表 3　各个模型响应时间

Table 3　Response time of each model

模型	训练方法	花费时间/s
RF	least	3.3
RF	random	1.8
MNB	least	1.7
MNB	random	0.5
LR	least	4.9
LR	random	3.9
GB	least	145.6
GB	random	143.2
SVM	least	676.8
SVM	random	679.3
BERT	least	7.2
BERT	DAL	12.5
BERT	DBAL	10.3
BERT	random	6.8

梯度提升树（GB）和支持向量机分类器（SVM）有着远超其他方法的时间消耗（分别为 2 分钟以上和 11 分钟以上），不适合作为主动学习过程中的任务模型。分析效率低的原因，支持向量机由于使用数据集的核矩阵（Kernel Matrix）描述样本之间的相似性，矩阵元素的个数随着数据规模增大成平方增长。当处理 TF-IDF 文本表示的 1 000 个维度的数据表示且训练样本量达到一定规模时，模型训练速度就会明显变慢。而梯度提升树分类器的弱学习器之间存在依赖关系，难以并行训练数据，同样难以处理大规模数据。

BERT 模型与深度主动学习方法因为能够利用 GPU 计算加速计算过程，所以速度虽然次于随机森林等模型在 5 秒内的响应时间，但其 10 秒左右的响应时间也不会让标注人进入等待，符合深度主动学习过程对模型的响应速度要求。

对比主动学习和非主动学习过程的模型响应时间，发现深度主动学习或者主动学习方法的模型处理耗时一般略高于随机采样，这是因为主动学习的采样策略相比非主动学习的随机采样需要更多计算步骤，如不确定性采样需要计算未标记池中每个样本的预测概率。

综合考虑模型精度和模型响应时间，最终在机器学习模型中选择了随机森林分类器（RF），朴素贝叶斯分类器（MNB）和逻辑回归分类器（LR）作为主动学习的任务模型，和 BERT 模型的深度主动学习方法进行对比。

3.2　深度主动学习实验

本实验测试深度主动学习算法（DAL、DBAL 和最低置信度 3 种方法）搭配 BERT 模型在实际新闻分类筛选任务中的表现。为了对比 BERT 模型的效果，还使用了几种经典的机器学习模型的主动学习过程作为对比。根据上一节的预实验的模型选择结果选择了随机森林分类器（RF），朴素贝叶斯分类器（MNB）和逻辑回归分类器（LR）。对于每种机器学习模型，都使用了最低置信度的主动学习方法，并使用随机采样作为对照。

实验进行了 20 次迭代，共 30 次重复实验。在主动学习迭代中各模型的 F_1 分数提升情况如图 2 和表 4 所示。可以看出整体而言，训练相同模型的主动学习方法相较于非主动学习，能够实现更快的精度提升，

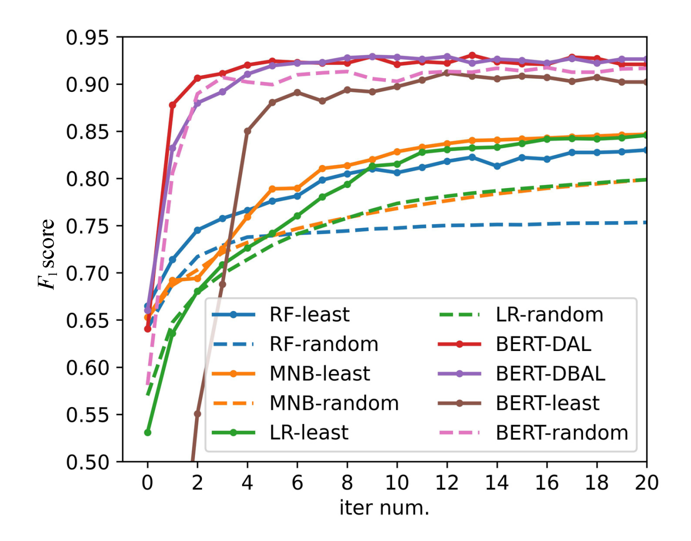


图 2　主动学习过程各模型 F_1 分数

Fig.2　F_1 score of each model in the process of active learning

这对于 BERT 模型和 3 种机器学习模型是一致的。

表 4　主动学习过程各模型 F_1 分数

Table 4　F_1 score of each model in the process of active learning

模型	训练方法	iter4	iter10	iter20
LR	least	0.838	0.940	0.975
LR	random	0.823	0.892	0.921
MNB	least	0.875	0.955	0.976
MNB	random	0.844	0.886	0.921
RF	least	0.883	0.929	0.957
RF	random	0.851	0.862	0.869
BERT	DAL	0.911	0.929	0.922
BERT	DBAL	0.892	0.929	0.926
BERT	least	0.688	0.891	0.902
BERT	random	0.901	0.906	0.916

将 BERT 模型的深度主动学习和机器学习模型的主动学习过程对比。可以看出 BERT 模型的 F_1 分数上升速度更快，都在 6 次迭代内就达到了最高值。而机器学习模型 F_1 分数上升速度较慢，而且一直落后于 BERT 模型。

对于 BERT 模型，表现最优的深度主动学习方法是 DAL 方法，而作为对比基线的随机采样方法则稍微低于 DAL 和 DBAL 两种方法。3 种方法在前 3 次迭代中 F_1 分数快速上升，而在 4~6 次迭代中已经非常平稳，总体呈现对数型增长。对于 BERT 模型表现最差的为最低置信度方法，其通过 6 次迭代才最终达到了其他方法在第 3 次迭代的性能，整体呈现均匀上升趋势。可能是最低置信度方法中 BERT 模型最后部分 softmax 层的输出值并不适合作为模型的不确定性度量。这导致其性能提升不仅慢于其他两种主动学习方法，还慢于随机采样方法。所以在后续新闻文本分类的实践场景中，应避免采用最低置信度方法对 BERT 模型就行主动学习训练。

总体来说，实验验证了在实际的农业新闻文本筛选任务中 BERT 模型配合深度主动学习方法的可用性和高效性，具体推荐使用 BERT 任务模型搭配 DAL（其次是 DBAL）采样函数作为深度主动学习方法。

4　结果讨论

4.1　深度主动学习选择策略分析

对不同的 AL 采样策略所获得的样本使用多样性指标和代表性指标进行比较，从而了解每种策略的特点为以后 AL 策略的选择与改进提供启发。

多样性：每次 AL 选择中，一批彼此之间较为不同的样本通常比选择一批相互相似甚至重复的例子更有效果。根据 ZHDANOV 的研究[26]，集合 B 的多样性可定义为：

$$D(B)=\left(\frac{1}{|U|}\sum_{x_i\in U}\min_{x_j\in B} d\left(x_i,x_j\right)\right)^{-1} \tag{3}$$

其中，x_i 表示用 L 训练的模型得到的示例 i 的[CLS]标记的表示，$d(x_i，x_j)$表示 x_i 和 x_j 之间的欧氏距离。

代表性：AL 策略（尤其是是基于不确定性的策略）的一个已知问题是它们倾向于选择不能正确代表总体数据分布的离群例子。因此，检查样本代表性能够检查是否存在该问题。本文使用 ZHU 等提出的 KNN- 密度度量[27]。其中一个样本的密度通过所讨论的样本集合中和它的最相似的 K 个例子的[CLS]表示在 U 内之间的平均距离来量化，而根据经验一般样本密度越高则越具有代表性。

图 3 描述了不同采样策略在对 BERT 模型的每轮训练中选择出的样本的多样性和代表性评估结果。我们对多次重复实验的结果取平均值，然后统计每步迭代上的指标均值和方差分布，从而得到指标值分布的箱线图。

在多样性指标上，旨在增加多样性的 DAL 方法和核心集方法具有最多样化的数据批次，并且 DAL 达到最高的多样性值。相比之下，其他策略倾向于选择选择较少多样性的数据。因此，将这些方法与强调多样性的方法相结合[26,28]可能会进一步提高其预测性能的结果。最低置信度方法的多样性又低于 DBAL 方法，这部分解释了对 BERT 模型训练时最低置信度方法性能提升过慢的原因。

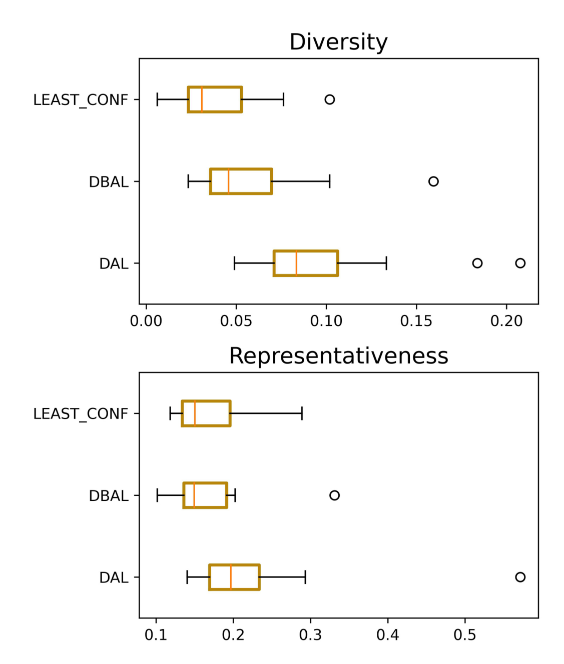


图 3 训练 BERT 时采样策略多样性和代表性评估

Fig.3 Evaluation of diversity and representativeness of sampling strategies

在代表性方面，DAL 作为一种代表性驱动的方法，同样在整个场景中始终领先。再考虑到 DAL 方法在 3 个实验中都表现除了稳定且优秀的性能，所以推荐在以后的新闻文本分类的 BERT 模型或者类似的 Transformer 架构模型的主动学习训练上首选该方法。其他两种主动方法的代表性分数则相互差别不明显。

最低置信度方法具有最低的多样性值，并且其代表性值也很低，这表明最低置信度这种简单的不确定性度量并不适合于深度网络。所以在实际应用时应避免使用该方法，或者将其作为深度模型主动学习实验中的一个基线对照组。

4.2 标注成本节约情况分析

对比试验中同一个模型的主动学习方法和非主动学习方法下达到相同 F_1 分数所需要的迭代次数（也就是数据标注数量），就可以分析深度主动学习或者主动学习方法所节约的数据标注的数量和比例。以非主动学习方法最终轮次的 F_1 分数的不同百分比划定不同的 F_1 分数标准，标注成本节约比例如表 5 和图 4 所示。

标注成本节约比例结果中最显著的特点是：以越高的 F_1 分数为标准对比主动学习与非主动学习的标注成本，主动学习方法的标注节约比例越高。所以代表 3 个模型不同标准下标注节约比例的 3 组柱状图内都在组内是从左到右逐渐增高的。分析原因是模型性能在随机采样中的提升过程是越来越慢的，当前模型 F_1 分数越高，进一步提升所花费的时间也越高。而主动学习过程在模型 F_1 分数越高时，对于训练过程的优化就越明显，能够更显著地提高训练效率。

表 5 各模型在不同标准下节约标注比例

Table 5 Each model saves annotation proportion under different standards

以最终 F_1 的 x%为标准/%	RF	MNB	LR
97	0.50	0.63	0.53
98	0.67	0.70	0.64
99	0.75	0.74	0.71
100	0.94	0.82	0.78

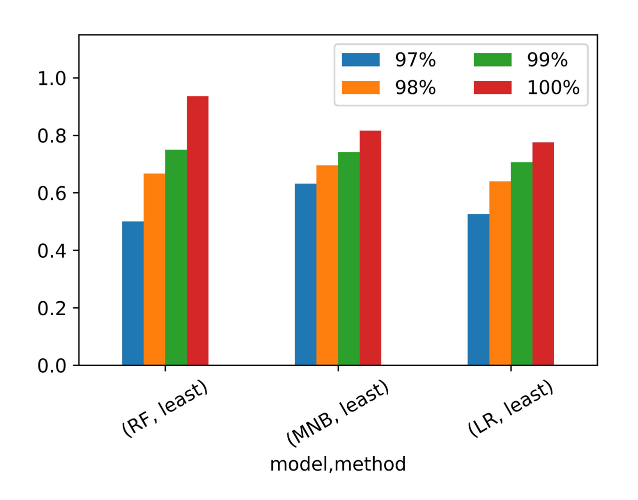


图 4 各模型在不同标准下节约标注比例

Fig.4 Each model saves annotation proportion under different standards

横向分析相对于各性能需求下不同模型的主动学习方法标注节约比例，可以看出在 99%和 100%被动学习最终 F_1 分数两个标准下，节约比例最高的模型是随机森林分类器（RF），分别节约非主动学习所需标注数据的 0.75 倍和 0.94 倍，其次是多项式朴素贝叶斯分类器（MNB）节约 0.74 倍和 0.82 倍，最后是逻辑回归分类器（LR）节约 0.71 倍和 0.78 倍。但是在 97%和 98%最终 F_1 分数两个标准下，随机森林分类器的标注节约比例相较于其他两个模型不再有优势。

因为 BERT 模型的提升迭代主要集中在前 5 次，采样点不够密集所以无法精确计算标注节约比例。但 BERT 模型的 F_1 分数提升过程同样是先快后慢的，推测也会有模型精度要求越高，标注成本节约比例越高的规律。例如 DAL 方法在第 3 个轮次达到随机采样在第 5 个轮次的 F_1 分数 0.902，但在第 6 次就达到了随机采样在第 16 个轮次的 F_1 分数 0.917。

5 结论与展望

本研究在 BERT 深度学习模型以及多种机器学习文本分类模型上，以爬虫收集的中文新闻数据为材料，以筛选出农业领域新闻为实验目标，验证了 3 种针对 BERT 深度网络的采样策略与任务模型配合后的主动学习效果，为新闻文本分类的深度主动学习提供了一种可操作可借鉴的实践经验。并将文本分类常用的机器学习模型，如随机森林分类器、多项式朴素贝叶斯分类器、逻辑回归分类器等结合最低置信度的主动学习方法分析与 BERT 模型对比分析。

实验证明，主动学习方法加快了深度模型的训练过程，并显著提高了其分类效果[29]。尤其是 BERT 模型配合 DAL 采样函数，是新闻文本主题分类与筛选场景下最佳的主动学习方案。其次可以选择 BERT 模型配合 DBAL 采样函数作为备选方案。在实验时还可设置随机采样作为基线对照方案。如果计算资源不足无法顺利训练 BERT 模型或者标注数据和标注资源较为充裕，则可以选择随机森林分类器等机器学习模型搭配最低置信度采样的主动学习方法作为替代方案。

通过记录每轮主动查询获取数据的多样性和代表性度量，尝试解释了不同采样策略的差异来源。发现 DAL 的多样性和代表性不仅强于随机采样（也即没有使用主动学习的情况），也强于其他两种主动学习方法，这部分解释了 DAL 搭配 BERT 模型在实验中为何表现最优。

在现有的爬虫数据集上，BERT 模型训练的主动学习方法和随机采样方法都在经过几次主动学习迭代后很快就达到了很高的分类性能，说明该数据虽然数量大但多样性却稍有不足，后续可考虑继续丰富新闻文本爬虫数据，进一步验证本研究方案可行性。

参考文献：

[1] 许丽, 焦博, 赵章瑞. 基于 TF-IDF 的加权朴素贝叶斯新闻文本分类算法[J]. 网络安全技术与应用, 2021, 11: 31-33.
XU L, JIAO B, ZHAO Z R. Weighted naive bayesian news text classification algorithm based on TF-IDF[J]. Network security technology & application, 2021, 11: 31-33.

[2] 郭文强, 李嫔. 基于 SVM 的新冠疫情虚假新闻检测[J]. 佛山科学技术学院学报(自然科学版), 2021, 39(6): 19-26.
GUO W Q, LI P. False news detection in the background of COVID-19 based on SVM[J]. Journal of Foshan university (natural science edition), 2021, 39(6): 19-26.

[3] 田沛霖, 符海滕, 马力禹, 等. 融合对抗训练和 CNN-BiGRU 神经网络的新闻文本分类模型[J]. 图书情报导刊, 2021, 6(8): 38-45.
TIAN P L, FU H T, MA L Y, et al. News text classification model based on adversarial training and CNN-BiGRU neural network[J]. Journal of library and information science, 2021, 6(8): 38-45.

[4] 刘子昂, 蒋雪, 伍冬睿. 基于池的无监督线性回归主动学习[J]. 自动化学报, 2021, 47(12): 2771-2783.
LIU Z A, JIANG X, WU D R. Unsupervised pool-based active learning for linear regression[J]. Acta automatica sinica, 2021, 47(12): 2771-2783.

[5] 黄永毅, 龚垒. 基于主动学习的交互式支持向量机文本分类学习方法[J]. 电子技术与软件工程, 2016, 14(14): 168-168.
HUANG Y Y, GONG L. Interactive support vector machine text classification learning method based on active learning[J]. Electronic technology & software engineering, 2016, 14(14): 168-168.

[6] 邱宁佳, 丛琳, 周思丞, 等. 结合改进主动学习的 SVD-CNN 弹幕文本分类算法[J]. 计算机应用, 2019, 39(3): 644-650.
QIU N J, CONG L, ZHOU S C, et al. SVD-CNN barrage text classification algorithm combined with improved active learning[J]. Journal of computer applications, 2019, 39(3): 644-650.

[7] 张智雄, 刘欢, 于改红. 构建基于科技文献知识的人工智能引擎[J]. 农业图书情报学报, 2021, 33(1): 17-31.
ZHANG Z X, LIU H, YU G H. Building an artificial intelligence

engine based on scientific and technological literature knowledge[J]. Journal of library and information science in agriculture, 2021, 33(1): 17–31.

[8] SENER O, SAVARESE S. Active learning for convolutional neural networks: A core-set approach[J]. Stat, 2018, 1050(2): 21.

[9] GAL Y, GHAHRAMANI Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning[C]. International conference on machine learning, 2016: 1050–1059.

[10] 杨承文, 李吉明, 杨东勇. 基于深度贝叶斯主动学习的高光谱图像分类[J]. 计算机工程与应用, 2019, 55(18): 166–172.

YANG C W, LI J M, YANG D Y. Active learning for hyperspectral image classification with deep bayesian[J]. Computer engineering and applications, 2019, 55(18): 166–172.

[11] DOR L E, HALFON A, GERA A, et al. Active learning for BERT: An empirical study[C]. Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), 2020: 7949–7962.

[12] HONEY J, LYNCH C D, BURKE F, et al. Ready for practice? A study of confidence levels of final year dental students at cardiff university and university college cork[J]. European journal of dental education, 2011, 15(2): 98–103.

[13] BELUCH W H, GENEWEIN T, NüRNBERGER A, et al. The power of ensembles for active learning in image classification[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018: 9368–9377.

[14] 李涛, 郭渊博, 琚安康. 融合对抗主动学习的网络安全知识三元组抽取[J]. 通信学报, 2020, 41(10): 80–91.

LI T, GUO Y B, JU A K. Knowledge triple extraction in cybersecurity with adversarial active learning[J]. Journal on communications, 2020, 41(10): 80–91.

[15] 徐睿, 梁循, 齐金山, 等. 极限学习机前沿进展与趋势[J]. 计算机学报, 2019, 42(7): 1640–1670.

XU R, LIANG X, QI J S, et al. Advances and trends in extreme learning machine[J]. Chinese journal of computers, 2019, 42(7): 1640–1670.

[16] BIAU G, SCORNET E. A random forest guided tour [J]. Test, 2016, 25(2): 197–227.

[17] RISH I. An empirical study of the naive bayes classifier[C]. IJCAI 2001 workshop on empirical methods in artificial intelligence, 2001: 41–46.

[18] 赵春晖, 高冰, 赵晨. 基于支持向量机和逻辑回归的半监督空谱加权的高光谱图像分类[J]. 黑龙江大学工程学报, 2019, 10(4): 64–72.

ZHAO C H, GAO B, ZHAO C. Semi-supervised spectral-spatial weighted classification of hyperspectral image based on SVMSLR framework[J]. Journal of Heilongjiang hydraulic engineering college, 2019, 10(4): 64–72.

[19] FRIEDMAN J H. Greedy function approximation: A gradient boosting machine[J]. Annals of statistics, 2001 , 29(5): 1189–1232.

[20] NOBLE W S. What is a support vector machine? [J]. Nature biotechnology, 2006, 24(12): 1565–1567.

[21] RAMOS J. Using TF-IDF to determine word relevance in document queries [C]. Proceedings of the first instructional conference on machine learning, 2003: 29–48.

[22] HAN K, XIAO A, WU E, et al. Transformer in transformer [J]. Advances in neural information processing systems, 2021, 34 (2): 15908–15919.

[23] BADRINARAYANAN V, KENDALL A, CIPOLLA R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(12): 2481–2495.

[24] 鞠默然, 罗江宁, 王仲博, 等. 融合注意力机制的多尺度目标检测算法[J]. 光学学报, 2020, 40(13): 1315002.

JU M R, LUO J N, WANG Z B, et al. Multi-scale target detection algorithm based on attention mechanism[J]. Acta optica sinica, 2020, 40(13): 1315002.

[25] PRECHELT L. Early stopping – But when? [J]. Neural networks: Tricks of the trade: Springer, 1998(1524): 55–69.

[26] REN P, XIAO Y, CHANG X, et al. A survey of deep active learning[J]. ACM computing surveys(CSUR), 2021, 54(9): 1–40.

[27] XIAO T, CAO F, LI T, et al. KNN and re-ranking models for English patent mining at NTCIR-7[C]. NTCIR, 2008.

[28] ALBERT-WEISS D, OSMAN A. Interactive deep learning for shelf life prediction of muskmelons based on an active learning approach[J].

Sensors, 2022, 22(2): 414-422.

[29] 金瑛, 叶飒, 李洪磊. 基于 ResNet-50 深度卷积网络的果树病害智能诊断模型研究[J]. 农业图书情报学报, 2021, 33(4): 58-67.

JIN Y, YE S, LI H L. The intelligent diagnosis model of fruit tree disease based on ResNet-50[J]. Journal of library and information science in agriculture, 2021, 33(4): 58-67.

A Classification Method of Agricultural News Text Based on BERT and Deep Active Learning

SHI Yunlai[1], CUI Yunpeng[1*], DU Zhigang[2]

(1. Agricultural Information Institute of CAAS, Beijing 100081; 2. Zibo Digital Agricultural Rural Development Center, Zibo 255000)

Abstract: [Purpose/Significance] At present, most of the training models used in the research of news classification are non-active learning. There are common problems about these models, including data cannot be labeled immediately and the labeling cost is too high, which also hinders the analysis of agricultural news. Especially because of the explosive growth of news data in the network era, it is more difficult to label data, train supervised text classification models, and screen relevant news in the field of agriculture from diversified online news sources. In order to solve this problem, the most commonly used pool based active learning or deep active learning technique is used to select more valuable and representative data from unlabeled data for manual labeling, and construct labeled data sets to improve the efficiency and effect of news classification and agricultural news mining. [Method/Process] The commonly used machine learning models for text classification, such as random forest classifier, polynomial naive Bayes classifier and logistic regression classifier, were combined with the active learning method with the lowest confidence to analyze the effect, and the BERT model was combined with the three sampling strategies of discriminative active learning, deep Bayes active learning and lowest confidence for deep active learning training. On the news corpus of 19 847 samples crawled and cleaned by crawler technology from Sina and other news websites, aiming at screening agricultural related news from diversified news samples of various topics, the iterative experiment of adding 30 samples per round was tested to check the improvement effect of F_1 score under various method combinations with the increase of the number of annotation. In addition, the representativeness and diversity of the samples selected by the sampling function of each method in the deep active learning method of the BERT model were compared, so as to understand the characteristics of each strategy and provide inspiration for the selection and improvement of Al strategy in the future. In addition, this paper also analyzed how much labeling cost can be saved by using the proposed method. [Results/Conclusions] When comparing a variety of machine learning models, it is found that although the gradient boosting tree and support vector machine classifier have high accuracy, they are not suitable for active learning because of their low efficiency in text data processing of large-scale high-dimensional data. After combining other machine learning models and the BERT model and training text models with the corresponding active learning or deep active learning methods, it is found that the application of active learning method can significantly improve the training process of each model. Among them, the BERT model, combined with discriminative active learning sampling function, has the best news text classification effect and the lowest annotation data requirements. The representativeness and diversity of the samples selected by discriminative active learning sampling function are also the highest, which explains the source of the advantages of this method. It can also be found that for the same task model, the higher the accuracy of classification is required, and the active learning method can save more annotation cost than non-active learning.

Keywords: deep learning; agricultural news; text classification; BERT model; active learning